Black box variation Bayes

"Black box variational Bayes" (BBVI) is a set of techniques for quickly and *automatically* approximating Bayesian posteriors using optimization. I'll consider "mean field automatic differentiation variational inference" (ADVI).

Black box variation Bayes

"Black box variational Bayes" (BBVI) is a set of techniques for quickly and *automatically* approximating Bayesian posteriors using optimization. I'll consider "mean field automatic differentiation variational inference" (ADVI).

Let θ denote model parameters and y some data and the joint generating distribution be $\mathbb{P}(\theta, y) = \mathbb{P}(y|\theta)\,\mathbb{P}(\theta)$. Let $\mathbb{Q}(\theta|\eta)$ be a family of candidate approximate posteriors, here taken to be independent normals.

ADVI aims to find

$$\overset{*}{\eta} := \underset{\eta}{\mathrm{argmin}}\,\mathrm{KL}\left(\mathbb{Q}(\theta|\eta)\,||\mathbb{P}(\theta|y)\right) = \underset{\eta}{\mathrm{argmin}}\,\mathbb{E}_{\mathcal{N}(z)}\left[f(z|\eta)\right]$$

for a cleverly constructed, automatically-differentiable $\eta \mapsto f(z|\eta)$.

Black box variation Bayes

"Black box variational Bayes" (BBVI) is a set of techniques for quickly and *automatically* approximating Bayesian posteriors using optimization. I'll consider "mean field automatic differentiation variational inference" (ADVI).

Let θ denote model parameters and y some data and the joint generating distribution be $\mathbb{P}(\theta, y) = \mathbb{P}(y|\theta)\,\mathbb{P}(\theta)$. Let $\mathbb{Q}(\theta|\eta)$ be a family of candidate approximate posteriors, here taken to be independent normals.

ADVI aims to find

$$\overset{*}{\eta} := \underset{\eta}{\operatorname{argmin}}\, \mathrm{KL}\left(\mathbb{Q}(\theta|\eta)\,||\mathbb{P}(\theta|y)\right) = \underset{\eta}{\operatorname{argmin}}\, \mathbb{E}_{\mathcal{N}(z)}\left[f(z|\eta)\right]$$

for a cleverly constructed, automatically-differentiable $\eta \mapsto f(z|\eta)$.

Unfortunately, $\mathbb{E}_{\mathcal{N}(z)}\left[f(z|\eta)\right]$ is typically intractable. So ADVI uses stochastic gradient (SG). The leads to the following problems:

- You have to tune the step size carefully
- You can't assess convergence directly
- You can't compute sensitivity, so you can't use linear response covariances.

$\Rightarrow$ Optimization is slow and imprecise, and the posterior uncertainty is no good. Not so black box actually!

We propose a simple alternative to SG that resolves these problems (sometimes).

Optimization of intractable expectations

Suppse you want to minimize an objective function of the form

$$\overset{*}{\eta} := \operatorname*{argmin}_{\eta} \mathbb{E}_{\mathbb{P}(z)} \left[f(z|\eta) \right] := \operatorname*{argmin}_{\eta} \ell(\eta),$$

where $\mathbb{P}(z)$ is known, but the expectation is not available in closed form.

Optimization of intractable expectations

Suppse you want to minimize an objective function of the form

$$\overset{*}{\eta} := \underset{\eta}{\operatorname{argmin}}\, \mathbb{E}_{\mathbb{P}(z)}\left[f(z|\eta)\right] := \underset{\eta}{\operatorname{argmin}}\, \ell(\eta),$$

where $\mathbb{P}(z)$ is known, but the expectation is not available in closed form.

When does this happen?

- Black box variational inference
- Stochastic control (e.g. you have a factory, and supply and demand are random)

Optimization of intractable expectations

Suppse you want to minimize an objective function of the form

$$\overset{*}{\eta} := \underset{\eta}{\mathrm{argmin}}\, \mathbb{E}_{\mathbb{P}(z)}\left[f(z|\eta)\right] := \underset{\eta}{\mathrm{argmin}}\, \ell(\eta),$$

where $\mathbb{P}(z)$ is known, but the expectation is not available in closed form.

When does this happen?

- Black box variational inference
- Stochastic control (e.g. you have a factory, and supply and demand are random)

What can you do? There are two options, both using the Monte Carlo (MC) estimate

$$\hat{\ell}(\eta) := \frac{1}{N}\sum_{n=1}^{N} f(z_n|\eta) \approx \ell(\eta).$$

Optimization of intractable expectations

Suppse you want to minimize an objective function of the form

$$\overset{*}{\eta} := \underset{\eta}{\mathrm{argmin}}\, \mathbb{E}_{\mathbb{P}(z)} \left[f(z|\eta) \right] := \underset{\eta}{\mathrm{argmin}}\, \ell(\eta),$$

where $\mathbb{P}(z)$ is known, but the expectation is not available in closed form.

When does this happen?

- Black box variational inference
- Stochastic control (e.g. you have a factory, and supply and demand are random)

What can you do? There are two options, both using the Monte Carlo (MC) estimate

$$\hat{\ell}(\eta) := \frac{1}{N} \sum_{n=1}^{N} f(z_n|\eta) \approx \ell(\eta).$$

- Stochastic gradient (SG)
    - Update with $\eta^{i} = \eta^{i-1} - \rho \nabla_{\eta} \hat{\ell}(\eta)$ for some step size ρ (new z_n every step)
    - Approximately minimizes the exact objective

Optimization of intractable expectations

Suppse you want to minimize an objective function of the form

$$\overset{*}{\eta} := \underset{\eta}{\operatorname{argmin}} \mathbb{E}_{\mathbb{P}(z)} \left[f(z|\eta) \right] := \underset{\eta}{\operatorname{argmin}} \ell(\eta),$$

where $\mathbb{P}(z)$ is known, but the expectation is not available in closed form.

When does this happen?

- Black box variational inference
- Stochastic control (e.g. you have a factory, and supply and demand are random)

What can you do? There are two options, both using the Monte Carlo (MC) estimate

$$\hat{\ell}(\eta) := \frac{1}{N} \sum_{n=1}^{N} f(z_n|\eta) \approx \ell(\eta).$$

- Stochastic gradient (SG)
  - Update with $\eta^{i} = \eta^{i-1} - \rho \nabla_{\eta} \hat{\ell}(\eta)$ for some step size ρ (new z_n every step)
  - Approximately minimizes the exact objective
- Sample average approximation (SAA)
  - Find $\hat{\eta} := \operatorname{argmin}_{\eta} \hat{\ell}(\eta)$ for fixed z_n
  - Exactly minimizes approximate objective

Optimization of intractable expectations

Suppse you want to minimize an objective function of the form

$$\overset{*}{\eta} := \underset{\eta}{\operatorname{argmin}} \mathbb{E}_{\mathbb{P}(z)} \left[f(z|\eta) \right] := \underset{\eta}{\operatorname{argmin}} \, \ell(\eta),$$

where $\mathbb{P}(z)$ is known, but the expectation is not available in closed form.

When does this happen?

- Black box variational inference
- Stochastic control (e.g. you have a factory, and supply and demand are random)

What can you do? There are two options, both using the Monte Carlo (MC) estimate

$$\hat{\ell}(\eta) := \frac{1}{N} \sum_{n=1}^{N} f(z_n|\eta) \approx \ell(\eta).$$

- Stochastic gradient (SG)
  - Update with $\eta^{i} = \eta^{i-1} - \rho \nabla_{\eta} \hat{\ell}(\eta)$ for some step size ρ (new z_n every step)
  - Approximately minimizes the exact objective
- Sample average approximation (SAA)
  - Find $\hat{\eta} := \operatorname{argmin}_{\eta} \hat{\ell}(\eta)$ for fixed z_n
  - Exactly minimizes approximate objective

Which is better? **In general, it depends.**

As far as we can tell, the BBVI literature has only ever considered SG.

Optimization of intractable expectations

Sample average approximation (SAA)

- Find $\hat{\eta} := \mathrm{argmin}_{\eta} \hat{\ell}(\eta)$
- Fixed z_n for whole procedure
- Exactly minimizes approximate objective

Stochastic gradient (SG)

- $\eta^i = \eta^{i-1} - \rho \nabla_{\eta} \hat{\ell}(\eta)$
- New z_n every step
- Approximately minimizes the exact objective

Optimization of intractable expectations

Sample average approximation (SAA)

- Find $\hat{\eta} := \mathrm{argmin}_\eta \hat{\ell}(\eta)$
- Fixed z_n for whole procedure
- Exactly minimizes approximate objective

Advantages:

- Can use fast off-the-shelf second-order optimization (great for poorly-conditioned problems)
- Can evaluate the objective function exactly to check for convergence
- Can compute sensitivity (linear response covariances ⇒ more accurate posterior covariances for mean field approximations)

Stochastic gradient (SG)

- $\eta^i = \eta^{i-1} - \rho \nabla_\eta \hat{\ell}(\eta)$
- New z_n every step
- Approximately minimizes the exact objective

Optimization of intractable expectations

Sample average approximation (SAA)

- Find $\hat{\eta} := \mathrm{argmin}_\eta \hat{\ell}(\eta)$
- Fixed z_n for whole procedure
- Exactly minimizes approximate objective

Advantages:

- Can use fast off-the-shelf second-order optimization (great for poorly-conditioned problems)
- Can evaluate the objective function exactly to check for convergence
- Can compute sensitivity (linear response covariances $\Rightarrow$ more accurate posterior covariances for mean field approximations)

Stochastic gradient (SG)

- $\eta^i = \eta^{i-1} - \rho \nabla_\eta \hat{\ell}(\eta)$
- New z_n every step
- Approximately minimizes the exact objective

Advantages:

- Uses each draw z_n only once (for a single gradient step)

Optimization of intractable expectations

Sample average approximation (SAA)

- Find $\hat{\eta} := \mathrm{argmin}_{\eta} \hat{\ell}(\eta)$
- Fixed z_n for whole procedure
- Exactly minimizes approximate objective

Advantages:

- Can use fast off-the-shelf second-order optimization (great for poorly-conditioned problems)
- Can evaluate the objective function exactly to check for convergence
- Can compute sensitivity (linear response covariances $\Rightarrow$ more accurate posterior covariances for mean field approximations)

Stochastic gradient (SG)

- $\eta^i = \eta^{i-1} - \rho \nabla_{\eta} \hat{\ell}(\eta)$
- New z_n every step
- Approximately minimizes the exact objective

Advantages:

- Uses each draw z_n only once (for a single gradient step)

This is actually a big one! Because if $\eta \in \mathbb{R}^D$, in general, both SG and SAA have accuracy $(D/N)^{-1/2}$, where N is the *total* number of draws of z_n used.

SAA uses each draw at each step of optimization. SG uses each draw once.
$\Rightarrow$ In general, SG is much more efficient in high dimensions!

Optimization of intractable expectations

Sample average approximation (SAA)

- Find $\hat{\eta} := \mathrm{argmin}_{\eta} \hat{\ell}(\eta)$
- Fixed z_n for whole procedure
- Exactly minimizes approximate objective

Advantages:

- Can use fast off-the-shelf second-order optimization (great for poorly-conditioned problems)
- Can evaluate the objective function exactly to check for convergence
- Can compute sensitivity (linear response covariances $\Rightarrow$ more accurate posterior covariances for mean field approximations)

Stochastic gradient (SG)

- $\eta^i = \eta^{i-1} - \rho \nabla_{\eta} \hat{\ell}(\eta)$
- New z_n every step
- Approximately minimizes the exact objective

Advantages:

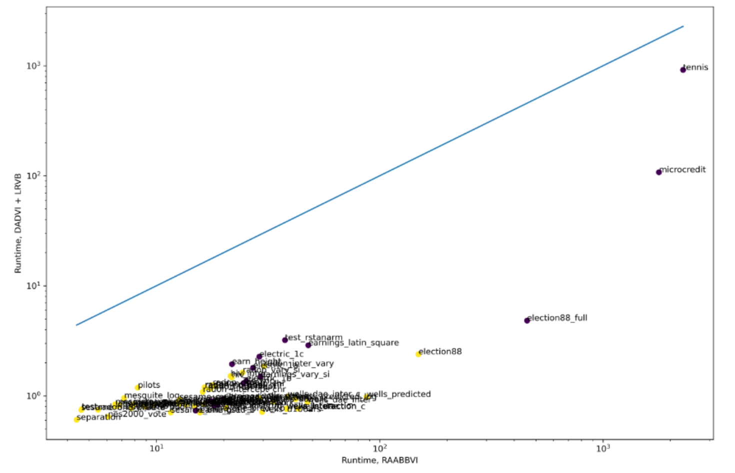
- Uses each draw z_n only once (for a single gradient step)

This is actually a big one! Because if $\eta \in \mathbb{R}^D$, in general, both SG and SAA have accuracy $(D/N)^{-1/2}$, where N is the *total* number of draws of z_n used.

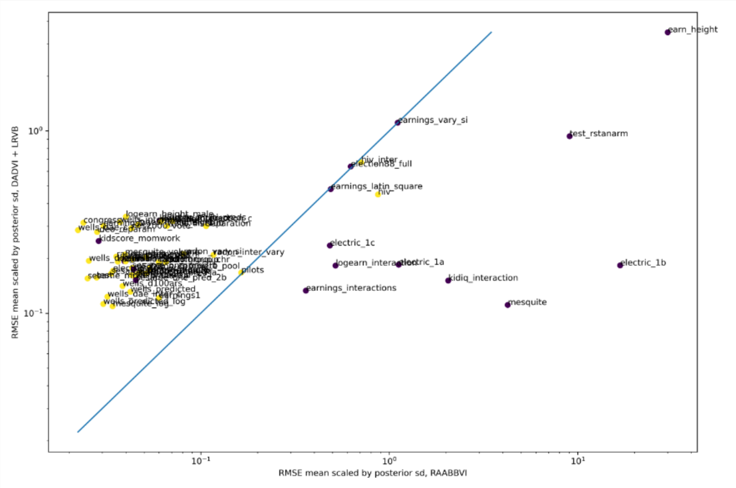
SAA uses each draw at each step of optimization. SG uses each draw once. $\Rightarrow$ In general, SG is much more efficient in high dimensions!

Theorem (us). If $\log \mathbb{P}(\theta, y)$ is high dimensional due to a large number of "local" variables, then the accuracy is $(\log D/N)^{-1/2}$, rendering SAA feasible.

Experimental results: Runtime



Experimental results: Means



Experimental results: Standard deviations